远程沉浸式自然交互

黄美玉 尹苓琳 纪雯 张博宁 王向东 陈益强

摘要：本文提出了沉浸式视频自然交互的概念和应用场景，同时介绍了几种能够加强沉浸式自然交互的方法，研究其中的关键技术。具体内容包括：通过准确的视频对话人提取和多个视频对话人的统一虚拟场景融合，实现不同空间位置的远程视频交互，使参与者犹如身处同一个虚拟会议室，以排除空间隔离感，增强沉浸式体验；通过矫正用户视频镜头中的头部姿态和视线方向，实现远程视频交户中的自然眼神交流；通过 QoE 模型研究远程视频交互时视频传输的自适应调整问题，保证在带宽不对称、终端不一致的情况下的高质量远程视频交互；通过采用远距离语音采集，支持多人任意参与远程语音互动，实现清晰连续、具有方向和距离感的高保真沉浸式音频交互。

关键字：沉浸式 视线矫正 抠像 QoE 模型 高保真音频

1 引言

以高速交互的视音频应用为基础、以人和家庭为中心的物联网电视正走进人们的生活。沉浸式视频交互以自然的操控方式将远方世界更逼真地拉入眼前，并提供信息服务。远程沉浸使分布在不同地点的使用者能够在同一虚拟空间协同工作，创造出“比亲自到对方现场还要好”的环境。沉浸式视频交互将以一种新型的交互方式满足现代人对快节奏、个性化、高质量生活的追求。

未来视频会议将会结合音频、视频、投影、通信等一系列技术来加强视频沉浸感，突显交互的自然性。当人们进入视频会议房间时，无需在固定位置开启视频设备建立连接，随时随地便可与对方进行远程视频交互。摄像头如同对参与者随时跟踪，将多角度可调节画面传送给对方。同时，传送过来的数据通过投影技术、虚拟场景融合技术，将对方人像立体呈现在参与者身边。音频技术将收集处理参与者音频信号。参与者无论处于房间何处，与对方的交流都如同带上了耳麦那样清晰。未来人们可以在房间内并发地进行自己的工作和参加视频会议。视频会议将以一种“面对面”、无干扰的自然方式渗入人们的生活。总之，这种所谓“沉浸式自然交互”指的是在不打扰用户原始活动情况下，让用户在远程视频交互过程中有同处一室的沉浸感，如同面对面一般自然交流。

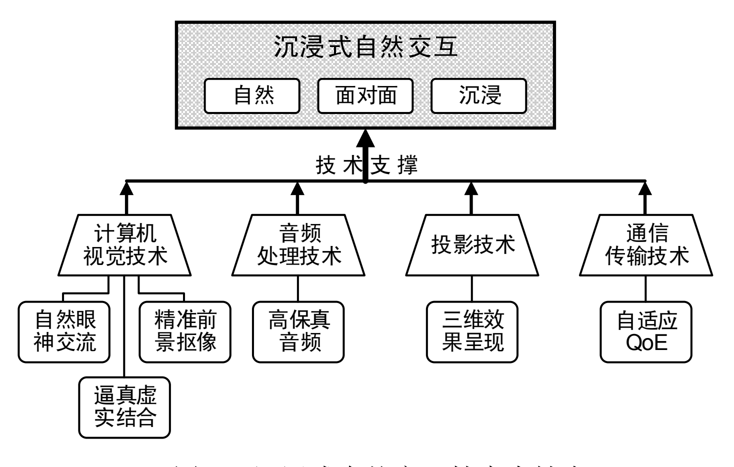


图1. 沉浸式自然交互技术支撑点

沉浸式视频是计算机视觉、音频领域极为重要又具有深远意义的研究课题，是未来多媒体视频技术的重要组成部分，具有极其广泛的应用前景。本文的主要工作是关于沉浸式视频中四个方面的研究（参见图 1）：

– **基于精准对象分割的虚实融合视频合成技术** 虚实融合是为了加强沉浸式视频会议交互者的临近感，为交互双方建立统一的虚拟场景，让交互双方有如同在一个地方的体验。这需要对人的精准抠像分割，对虚拟场景进行建模，并将抠像结果与虚拟场景进行融合。

– **基于深度伪三维信息合成的自然眼神交互技术** 眼神交互关系着交互的自然感。但在实际的视频交互中，人们往往不能同时凝视屏幕和摄像头，没有眼神接触、凝视感等效果，破坏了远程视频交互的沉浸感，必须通过眼神交互技术解决这一问题。

– **面向异构网络和终端的沉浸式用户体验质量自适应技术** 提高用户体验质量（QoE，Quality of Experience）是互联网的发展目标，远程沉浸应用也依赖于用户体验质量的支持。因此，远程沉浸视频必须建立服务质量控制机制，系统应能够可靠地评估和预测服务质量，以实现系统参数的自适应调整。

– **沉浸式高保真音频交互技术** 为营造自然、沉浸式的互动氛围，在音频方面，应该使音频采集透明化，即说话人完全无需在意音频采集设备的存在，可以像面对面交谈时一样随时自由发言；另一方面，由远程传来的音频在输出时应具备方向、距离等位置感，产生出身临其境的共同交谈的感觉。

2 国内外发展现状

2.1 基于精准对象分割的虚实融合视频合成技术

基于精准对象分割的虚实融合视频合成技术涉及的子技术主要包括：（1）在线视频精准对象分割技术；（2）虚拟场景建模技术；（3）虚实融合技术。

2.1.1 在线视频精准对象分割技术

在线视频精准对象分割是实现沉浸式视频融合的关键技术手段，也是最近几年才开始的一个比较新的研究方向。由于在线视频对象分割过程不能利用用户交互，且其对算法的速度和鲁棒性都有较高的要求，因此到目前为止还处于非常初级的研究阶段，只能处理摄像机固定，且背景相对静止的情况，速度上也只是勉强能达到应用的需求。现有虚拟演播室系统一般都是采用背景标记法的蓝屏抠像技术，就是在特定颜色（一般为蓝色）的场景拍摄前景对象视频，形成该颜色背景，这样利用图像处理的方法便可将前景对象提取出来。背景标记方法对背景有特定的要求，且摄像机无法得到前景对象的三维坐标，影响后期的三维场景视频合成效果与真实感。由于深度信息对于光照变化，动态阴影具有鲁棒性，使用深度信息可以改善前景提取的质量。 但是，通过深度传感器获得的深度信息在边界处容易产生错误，而通过立体匹配技术获取的深度信息容易在平坦区域产生错误。因此，需要采用融合技术实现两种深度信息的互补和矫正以实现精确的前景提取。另外也可以先基于深度传感器获取的深度信息实现前景的粗分割，然后采用彩色图像提取的信息来修正分割结果，这方面的技术有：

(1). **羽化算法** 该方法速度非常快，但只能处理轻微的错误，并且羽化之后原本清晰的边界会变得模糊，与背景合成之后会有明显的颜色溢出；

(2). **抠图算法** 现有常用的抠图方法有：剔除（Knockout）、泊松抠图(Poisson matting)、贝叶斯抠图(Bayes matting)、边界抠图（border matting） 等方法。在一些交互的视频前景提取系统中，边界抠图被扩展到视频体上。虽然这样有助于改善结果沿时间轴上的一致性，但在视频体上难以以此进行实时优化。此外，基于抠图的方

法通常都假设前、背景之间是光滑过渡的，因此也不可避免地会在边界清晰的地方造成背景色溢出。当前景和背景颜色相似时，这种错误会很严重。一致抠图（Coherence matting）在贝叶斯抠图的基础上施加了边界约束，这样可以增强算法在前景和背景颜色相似环境下的鲁棒性。但是抠图的方法运行速度较慢，很难满足在线视频分割的实时性要求。

(3). 后处理算法 该算法是一种基于颜色信息和边界信息的实时边界修正方法，能够自适应地调整过渡区域的宽度，使边界清晰但不生硬，从而能防止背景合成时的颜色溢出。但是该方法没有考虑到时序一致，不能消除视频中的闪烁问题。

2.1.2 虚拟场景建模技术

虚拟场景建模，即构造虚拟世界，是三维场景虚实融合技术中的另一重要问题。虚拟三维空间良好建模是产生沉浸感和真实感的先决条件。场景太简单，会使用户觉得虚假，而复杂逼真的场景又势必会增加交互的难度，并影响实时性。当前虚拟场景建模的方式主要有：基于图形渲染的建模技术、基于图像的建模技术和基于图形与图像的混合建模技术。

(1). 基于图形渲染的建模技术

基于图形渲染的建模方法是充分利用计算机图形学技术进行虚拟环境的建模和渲染。首先对真实世界进行抽象，建立数学模型(一般是几何多边形)，然后给定观察点和观察方向，利用计算机根据该模型实现多边形处理、着色、消隐、光照以及投影等一系列绘制过程，产生虚拟场景。因此，虚拟物体的几何建模、表面材质的纹理映射、视点光照的处理是基于图形渲染要解决的主要问题。

(2). 基于图像的建模技术

基于图像的建模技术的本质是图形学中广为应用的纹理映射技术，即用待建三维虚拟空间的有限幅图像样本，在一定的图像处理算法和视觉计算算法的基础上，来直接构造三维场景。基于图像建模的方法可以克服复杂场景三维建模的困难，并且可以使用真实世界的图像提供更丰富的细节，较容易得到与真实环境相近的效果，生成图像的质量独立于场景的复杂性。其计算量较小，也不受场景复杂度的限制，且对硬件的要求也不及基于图形的建模那样高，还可以在微机上实现。但由于场景中的虚拟物体是图像中的二维对象，因而用户很难，甚至不能与这些二维对象进行交互，出现漫游失真。因此，该方法仅适合于基于真实自然场景的三维虚拟环境的建立。

(3). 基于图形与图像的混合建模技术

既要避免复杂场景几何模型的大计算量，又要满足实时性要求，可以采用基于图形与图像的混合建模方法。在虚拟现实混合建模中，用户可以用“用户化身(User avatar)”这个特殊的虚拟实体对象的形式进入虚拟场景，即用户与虚拟场景的交互是通过用户化身与场景中其它虚拟实体对象间的数据交换来完成的。基于图像的建模技术注重虚拟场景的视觉真实性，可用于交互要求少并且场景复杂的场合，用图像的插补、变形、拼合等方法来构造一个尽可能符合视觉要求的纯虚场景。尽管纯虚场景中的虚拟物体是二维图像中的纯虚对象，用户化身不能与之交互，但人们仍可凭借深度传感器技术来获取用户化身相对于图像中纯虚对象的深度信息。基于图形渲染的建模技术注重虚拟场景交互行为的仿真和可实现性，可用于用户希望与之产生交互作用的场景对象。

2.1.3 虚实融合技术

它的实质是将计算机制作的虚拟三维场景与实时分割的前景对象进行数字化的实时合成，使人物与虚拟背景能够天衣无缝地融合，以获得完美的合成画面。逼真的虚实融合要求前景图像和虚拟场景图像的合成图像保持透视一致、几何一致和光照一致。为了实现图像合成的透视一致和几何一致，在现有的虚拟演播室系统中都将虚拟摄像头和真实拍摄的摄像头进行了对准，因此只需使用色键技术，即将前景图像和虚拟场景图像进行简单的叠加，便可以实现两者的一致融合。为了实现图像合成的光照一致，现有的为人们所熟知的方法是泊松法。泊松法是在梯度域内，基于泊松偏微分方程来进行图像合成。当片图像（合成图像的前景）和背景图像的亮度相差很大时，泊松法不能得到满意的合成图像，因为泊松法是将背景图像的信息表示为狄雷克利边界条件。当背景图像在合成位置颜色值很接近时，狄雷克利边界条件可近似为常数，此时泊松法能合成出逼真的图像；但是，当合成位置颜色值相差较大时，狄雷克利边界条件是一个非常数边界，因此泊松法效果不好。由于光照一致性反映在颜色协调上，因此很多研究者对颜色协调进行了研究。不同图像之间的颜色协调包括亮度和色度两个方面。事先定义一定数量的颜色协调模版是绘画广泛采用的技术。一些研究者把合成图像的颜色协调表示为图像与颜色协调模板的最佳匹配问题，通过调节图像的亮度、饱和度使之与颜色协调模板相一致。还有一些研究者提出了一种基于色度和亮度的颜色距离计算方法，能够量化视觉突出性。图像的视觉突出性是指图像内在的、不变的视觉特性，描述了场景对人视觉刺激敏感的程度、位置。人眼能够利用这种特性轻松地识别物体。

2.2 自然眼神交互技术

自然眼神交互技术是实现沉浸式远程视频交互的关键技术，在辨识视频中的轮流发言、知觉性的注意力和意图及其他方面信息的过程中都有很重要的作用。然而，在传统的远程视频交互系统中，摄像头一般置于视频屏幕上方，在本地和远程视频之间的双向眼神交互是不太可能的。这个矛盾来源于摄像头光轴和人注视屏幕视线之间的夹角。视频对话的双方由于要注视处于屏幕上的对方而无法正视摄像头，使得视频中对话人多呈低头而非平视姿态，无法进行眼神的交互。研究表明，当摄像头光轴和人注视屏幕视线之间的夹角大于 5° 时候，眼神交流的丢失就比较明显了，而传统的电话会议系统不支持眼神交流。

为了实现自然眼神交流，目前的解决办法主要分为两类。一类是通过改变硬件设备实现，另一类则是通过软件矫正的方式实现。在基于硬件的方法中，一种是通过安置多个摄像头，用 GPU 实现插值技术保证眼神交互；另一种则是将半透明反射镜呈一定角度对准视线凝视的位置，从而达到克服眼神缺失的目的。但基于硬件的方法需要昂贵的设备费用和复杂的配置，因此在现实中难以推广。

目前，基于计算机视觉和图像处理的方法在视线矫正中被广泛采用。现有的方法中一种是通过双摄像头进行立体分析，可以得到场景的深度图，进而对多幅图像进行融合，获得中心虚拟视角的效果，从而保证眼神交互。但是此方法的硬件需要按规定安置，并且需要人工参与摄像机外参数标定，硬件设备一旦固定就不能轻易变动。同时由于多幅图像的融合点被固定在多摄像机的中心位置，无法对用户的位置变化自适应，使得矫正后眼神依然有偏差。另一种方法是将眼睛虹膜检测、眼睛轮廓检测算法用于眼神矫正中，一旦眼睛的位置被确定，可通过变换眼睛局部图像来达到眼神直视的效果。但是算法没有针对身体和头部进行相应的处理，使得矫正之后的图像中眼神跟人身体其他部分不协调。

2.3 沉浸式视频的用户体验质量模型研究

沉浸式视频的用户体验质量模型研究，是视频领域研究学者热门研究课题，领域的许多研究者提出了用户体验质量的评价方法来自动处理视频质量的问题。在经过影像压缩之后，

输出的影像通常都会在某种程度上与原始影像不一样。为了衡量经过处理后的影像品质，通常会参考峰值信噪比（PSNR，Peak Signal to Noise Ratio）值来认定某个处理程序是否令人满意。由于峰值信噪比值只有当图像在接收端重新建立时才进行计算，所以它不适用于实时系统，这也是这种方法的一个弊端。

用户体验质量最初是从通信领域提出的概念，定义为用户可以感知的服务质量，即终端用户对于网络提供的通信业务性能的主观感受。随着媒体技术的发展，用户体验质量延伸的范围越来越广，通常通过对用户体验质量进行量化分析来研究用户对服务质量（QoS）的体验和感受。用户体验质量和服务质量参数都可以反映服务质量的好坏，但是用户体验质量主要强调了用户的主观感受，而应用层服务质量从客观性能参数方面来评价媒体的服务质量。通常应用层服务质量参数性能越高，则用户感受的用户体验质量也就越好。

以往的研究重点多在于服务质量的保证，但是最终衡量视频品质的标准在于用户体验质量。服务质量是为了保证或提高用户体验质量而应用在网络上的技术指标，与具体的业务相关联。相对于不同的业务而言，服务质量的标准不一定适用于用户体验质量；有限的资源不可能对每个业务的每个用户都保证同等的服务质量。

2.4 沉浸式高保真音频交互

当前已有的远程视音频互动系统多数面向视频会议的应用，在音频方面，一般为每位参与者单独配备麦克风，在参与者发言时开启麦克风以开始音频采集。这种方式可避免环境噪声的影响，提高音频质量，但是操作相对繁琐。另外，对于不同发言人，以完全相同的方式采用近讲麦克风采集其语音，因此远端听到的语音没有方向和距离感。这样的系统不大符合远程亲情互动对音频的要求。

对于远程亲情互动而言，首先，系统应简单易用，使说话人无需手持、佩戴或靠近麦克风和执行额外操作，并应支持多人自由交谈。

其次，为营造沉浸式的亲情互动氛围，输出音频应能重现方向、距离等位置感，产生出身临其境，共同交谈的感觉。这要利用心理声学中著名的哈斯效应“欺骗”人耳的方法，其作用是使人能产生音源的位置感觉。传统的双通道立体声是无法分出上方和下方的音源的，所以严格意义上来讲，并不能称之为三维音频效应，而且在这种播放系统中，只有位于中心点的最佳听音点的听众，才可以听到再生的空间效果。

3 我们的研究工作介绍

远程视频正向大屏幕、高分辨率方向发展。下一代的远程视频要求在自然沉浸式方面有所突破。我们在上述四个增强沉浸式的方向上提出了自己的解决方案，形成了沉浸式自然交互的技术突破。

3.1 基于精准对象分割的虚实融合视频合成技术研发

3.1.1 在线视频精准对象分割技术

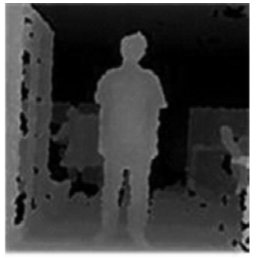
考虑到深度数据对天气、光照、阴影等因素的鲁棒性，本文采用深度摄像头与 RGB 摄像头的数据协同工作，解决复杂场景下人物对象的分割。图 2 中，（1）为普通摄像头获取的 RGB 图像，（2）为深度摄像头同步采集的深度图像。

在我们的研究中，首先采用深度摄像头的实时深度图像识别出人体对象的三维坐标及区

域，然后将深度图像的人体对象区域映射到同步采集的 RGB 图像上，实现人体对象在彩色图像上的分割。图 3 （1）为在深度图像中人体对象分割结果，将深度图像中人体对象分割结果直接映射到 RGB 图像中，可实现人体对象与背景的分离，如图 3（2）所示：

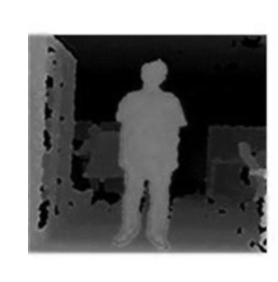


（1）RGB 图像



（2）深度图像

图2. 同步采集的 RGB 图像和深度图像



（1）基于深度图像的人体分割结果



（2）RGB 图像的人体分割结果映射

图3. 人体对象分割效果图

从图中可以看出，直接将深度图像的人体对象分割结果映射到 RGB 图像中得到的分割效果比较差，最主要原因在于由深度传感器获得的深度图像在深度非连续处容易出现错误或丢失，以致基于深度图像的人体对象分割容易在前景和背景的交界处出现误分割。我们采用了高效优质的后处理算法，以期实现精准人物对象分割。该后处理算法融合颜色、边界、运动和时序信息，以实现时序一致的边界优化效果。该算法首先使用局部颜色模型和边界函数计算每个待处理像素的局部 alpha 值[1]，然后采用一种简单的运动估计法，估算当前帧与相邻前两帧的运动概率图，接着以运动概率作为权值，求取局部 alpha 值和时序 alpha 值的加权和作为待处理像素的 alpha 值。由于上述的边界优化算法只能消除前景与背景交界处的误分割现象，却不能处理前景内部由于深度丢失造成的误判现象，当目标人佩戴眼镜或者头发披散在肩头的时候，在眼镜与脸部或者头发与肩部的交界处都会有深度丢失的现象，会形成前景误判为背景的孔洞，这些孔洞的出现将会大大影响分割的精度。因此需要将这些孔洞进行前景填充。然而，由于目标人的叉腰动作也会在前景内部形成孔洞，所以不能将前景内部的所有孔洞进行笼统地填充。本文根据深度传感器获得的深度数据的特性，给出了一种前景孔洞的判别算法。

首先，我们通过轮廓算法找到二值分割图像中的所有轮廓，然后判断每个轮廓的内部区域是否存在深度不为 0 的像素。如果存在，则该轮廓内部区域不需要进行填充，因为该轮廓不是由于深度缺失造成的孔洞的边界。当轮廓包含区域的像素的深度全部丢失的时候，该轮廓所包含的区域成为候选前景孔洞。对于候选前景孔洞，本文给出一种基于轮廓边缘和区域背景颜色相似度的前景孔洞判别算法。区别于传统的梯度边缘，本文采用边界清晰度作为判别轮廓像素的边缘的依据。另外，为了快速对动态背景建模，本文采用累积背景直方图对背景像素的颜色分布建模。









图4. 实时前景提取效果图

算法效果如图 4 所示。

[1] 颜色中的第四个成分，通常 alpha 值为 1 表示完全不透明，0 表示完全透明

3.1.2 虚拟场景建模

沉浸式视频需要与用户的行为产生交互，实现沉浸感。因此本文采用便于用户交互的基于图形渲染的建模技术，并应用已有的建模技术和纹理映射技术以及阴影处理技术给出一套建模方案。另外，为了增强虚拟环境的沉浸性，本文还将给出一套基于用户位置感知的互动交互方案。

模型的建立是创建虚拟环境的基础。对于需要实时与用户进行互动交互的虚拟环境，对模型的处理显得尤为重要。过多的模型细节会严重降低模拟的速率，为了在模型的细节和复杂性之间寻求平衡，我们使用纹理代替部分模型细节。另外，我们在建模过程中对模型进行分块，这样既可以减小建模的难度，又可以通过分块显示提高仿真效率。在实际的建模过程中，本文根据需建模型的特点选择建模方法。对于有规则平面的几何体，采用多边形建模方法；对于复杂曲面的几何体，则使用面片或者 NURBS[2]。因为要达到同样的曲面效果，面片和 NURBS 需要的节点和面数要少些。对于纹理映射问题，本文主要采用最近广泛应用的环境纹理映射技术实现材质的逼真纹理映射，对于明暗处理，则主要使用 Phong 模型[3]。

另外由于我们采用了深度传感器，因此可以精确地感知用户的位置信息，从而可以实现虚拟场景与用户的互动交互，增强用户的行为沉浸感。

3.1.3 虚实融合

在本文中，前景图像来源于远端摄像头拍摄的真实画面，虚拟场景图像来源于本地虚拟摄像头实时绘制的虚拟画面，为了实现两者的无缝融合，我们的研究采用了如下方案：

(1). 基于前景姿态的环境一致的图像合成

在本文中，针对坐和站两种姿态构造了不同的虚拟场景，以便实现环境一致的图像合成。在远程视频呈现的客户端，通过网络接收到填充的视频图像后，首先根据填充方案，进行二次抠像，还原远程人物对象的图像，然后根据前景的轮廓，辨别前景的姿态并据其选择虚拟场景，最终将前景叠加到虚拟场景的合适位置，实现逼真的合成效果。对于站姿构造的虚拟场景，我们将使其可移动空间最大化，避免出现前景穿过虚拟场景物体的失真现象。



图5. 站姿合成效果图　　图6. 坐姿合成效果图

(2). 基于深度信息的几何一致的图像合成

由于在本文中，拍摄前景图像的远端摄像头和绘制虚拟场景的虚拟摄像头并不一定存在校准关系，因此，不能直接将前景图像叠加到虚拟场景图像中，而应该重新计算前景图像在

[2] Non-Uniform Rational B-Splines，非均匀有理 B 样条曲线。一种建模方式

[3] 一种光照模型，可以表述为：由物体表面上一点 ***P*** 反射到视点的光强 ***I*** 为环境光的反射光强 $\boldsymbol{I_e}$、理想漫反射光强 $\boldsymbol{I_d}$ 和镜面反射光 $\boldsymbol{I_s}$ 的总和

虚拟场景中的缩放比例，实现几何一致的图像合成。

本文中，由于采用了深度摄像头，因此可以根据深度信息快速获取前景图像中每个像素点的真实三维坐标，从而获取前景图像在真实世界的高度。假设 ***H*** 为远端前景图像在真实世界的高度，***N*** 为远端前景图像在远端视频图像中的高度像素数，则远端前景图像在本地虚拟场景图像中的缩放比例可以通过下式计算：

$$S=d\times H/N$$

其中，***d*** 是本地虚拟场景中叠加位置处的“像素分辨率”（单位长度的像素数）。计算出缩放比例后，便可以通过对应的放大或缩小操作实现远端前景图像和虚拟场景大小一致的逼真融合。

(3). 基于视觉突出性的光照一致的图像合成

我们针对图像合成中的光照一致问题，采用一种保持视觉突出性的光照一致图像合成方法。该方法由视觉突出性计算、白点校正和根据视觉特性进行光照亮度调整三部分组成。视觉突出性通过综合色度和亮度特性计算得到；白点校正通过对齐片图象和背景图像的颜色主轴使片图像和背景图像达到白点一致；基于视觉突出性的光照亮度调整把光照一致图像合成表示为受视觉突出性约束的非线性优化问题，从而使合成图像不仅具有光照一致特性，且能保持原来的视觉突出性。

3.2 基于深度伪 3D 信息合成的自然眼神交互技术研发

针对现有眼神矫正方法的不足，我们提出一种基于虚拟视角的眼神矫正方法。与传统方法中对多摄像头获取的图像在固定虚拟视角点进行融合的技术路径不同，我们设置的虚拟视角将跟随人眼的位置变动。具体做法是将真实摄像机下获得的三维数据转换为虚拟摄像机坐标系下的三维数据，再重投影到虚拟摄像机成像平面。同时，现有方法中一般需要对摄相机模型中的外参数进行人工标定，而且要求硬件设备不能随意变动，因此降低了方法的灵活性。针对此问题，我们提出了基于虚拟坐标系的几何模型。利用该模型能对硬件距离虚拟坐标系原点的偏移量进行自标定。这使得在对摄像机内参数进行一次性标定之后无需用户参与外参数标定过程，也无需固定硬件设备位置，增加了普适性。方法流程如图 7 所示

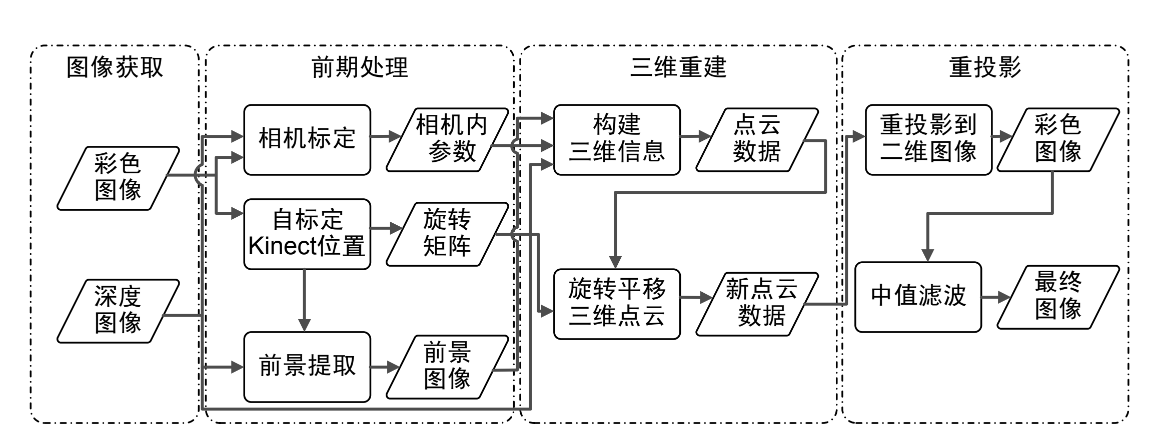


图7. 漂移量自标定方法流程图

首先，该方法需要获取彩色图像及与彩色图像对齐的深度图像作为原始数据。在前期处理阶段使用传统方法对摄像机内参数进行一次标定。由于虚拟坐标系下的外参数会因摄像机的位置发生改变而变动，我们提出了一种在前期处理阶段对外参数进行自标定的方法。根据摄像机内参数和图像深度数据依据摄像机模型即可获得摄像机坐标系下的三维点云数据；再

根据虚拟坐标系下的摄像机外参数，将此点云数据转换为虚拟坐标系下的点云数据，最后再次利用摄像机内参数将虚拟坐标系下的点云数据重投影到二维虚拟成像平面上，从而达到视线矫正的目的。

在虚拟摄像机的摄像机模型中，我们将真实世界坐标系设为彩色摄像头的坐标系。虚拟摄相机的摄像机坐标系即为虚拟坐标系，几何模型如图 8 所示。算法将左眼睛中心位置平行凝视摄像机平面的点设为虚拟视角点也即为虚拟坐标系的原点。彩色摄像头距离虚拟视角点的水平和垂直偏移量分别为 X_{off} 、Y_{off}。彩色摄像头仰视人眼的角度和水平偏移人眼的角度设为 θ′，α′。虚拟坐标系决定了虚拟照相机的安放位置。为了消除仰视和水平方向的偏移感，本文将虚拟坐标系进行了垂直和水平方向的角度旋转。虚拟视角坐标系 Y′ 轴和 Z′ 轴为彩色摄像头坐标系的 Y 轴和 Z 轴绕彩色摄像头坐标系的 X 轴方向旋转了 θ′，即摄像机仰视人眼的角度，这是为了矫正头部和视线的仰视。进而，虚拟视角坐标系的 X 轴和 Z′ 轴绕虚拟坐标系的 Y′ 轴再旋转 α′ 形成 X′ 轴和新的 Z′ 轴，即摄像机水平偏移人眼的角度，这是为了修正水平方向带来的偏移感。如此就形成新的虚拟坐标系如上所述，本文假设虚拟相机内参数与彩色摄像头内参数保持一致，而对应于不同位置下的硬件设备，虚拟摄相机外参数是不同的，因此算法需要根据上述提出的几何模型进行外参数自标定。

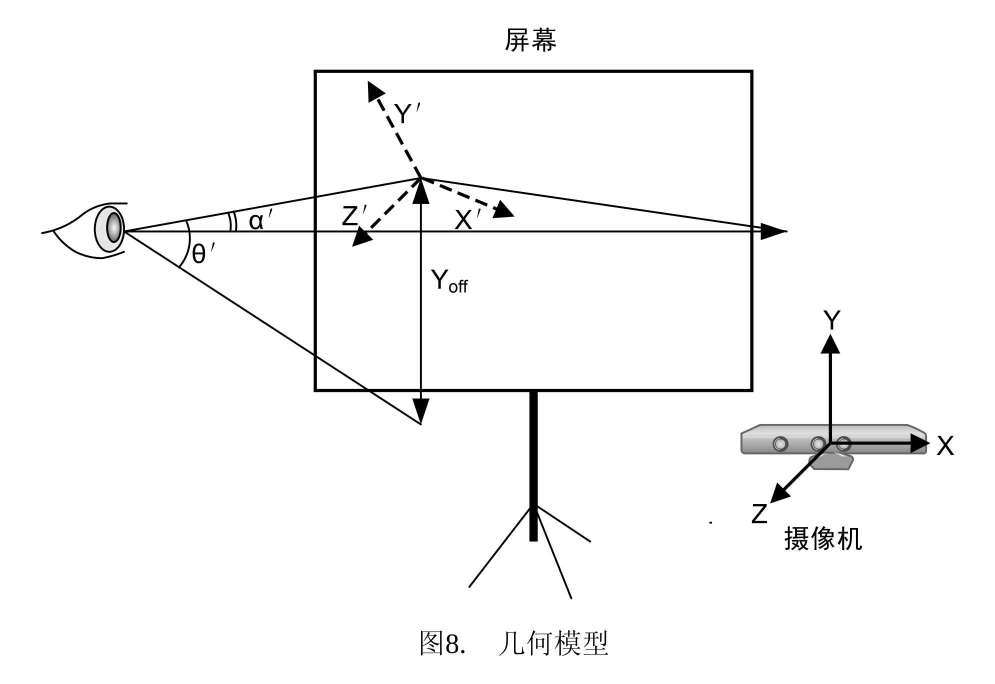


图8. 几何模型

通过虚拟摄像机外参数的自标定，对三维数据进行旋转变换，再利用摄像机成像模型映射到二维图像上，就得到最终的效果图。

3.3 面向异构网络和终端的沉浸式用户体验质量自适应技术研发

本文对建立面向异构终端的用户体验质量模型并实现其自适应，以构筑多通道自适应视频传输提出了以下的研究方案。

3.3.1 异构终端的沉浸式视频用户体验质量模型

我们采用实验和理论相结合的研究方法。既有实验部分对用户体验质量的分析，也有理论部分对构造多媒体业务性能衡量体系的映射模型建立。下面对用户体验质量参数实验研究部分以及用户体验质量和服务质量参数之间的映射关系部分分别进行阐述，并最终根据用户体验质量和服务质量参数之间建立的映射关系构建了端到端的多媒体业务质量衡量体系。

测取用户体验质量的数据一般分为主观和客观两种方法。现在的客观评价主要是基于均方误差或者峰值信噪比，这些指标有明确的物理意义，测试比较容易，但是由于没有考虑到人类视觉特性（HVS）,所以评估结果与主观感受并不相符。特别是，当可伸缩视频的帧速率，空间分辨率被联合调整后，计算出来的均方误差或者峰值信噪比总是不能反映视频的主观质量。

主观方法需要依靠对用户的调查和用户的投诉分析等方式，从用户体验的角度收集多种业务的用户体验质量统计数据。但是这样做局限性较大，不适合长期大规模的使用，只能在固定场合进行小规模的验证，用来检测客观类方法取得的数据的准确性并为客观方法提供参考依据。以下为具体的方法过程。

分析视频的特点
设计和确定QoE/QoS之间的关系
确定测试环境和测试样本
进行主观测试
分析数据的合理性
选择QoE模型

图9.　面向异构终端的用户体验质量模型建模过程

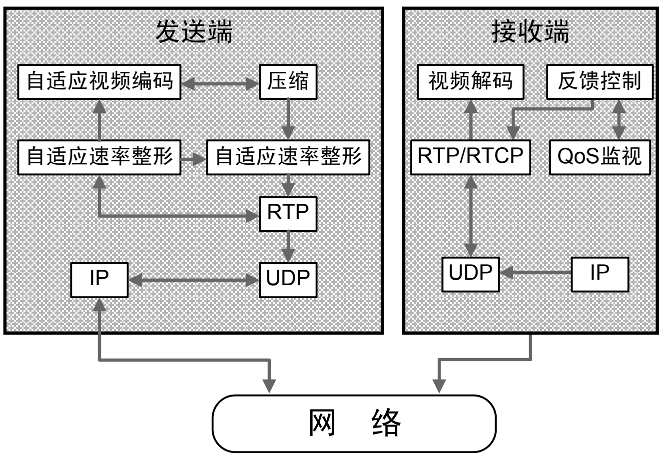
用户体验质量是完全从用户的角度出发测得的参数，我们通过引入切实可靠的实验数据来具体地分析这层参数。目前较广泛采用的是国际电信联盟（ITU）建议的“平均值估计法”（MOS），它将用户体验质量的主观感受分为 5 个等级，用此种量化的方法较为细致地描述了用户的主观感受。

我们的研究围绕用户体验质量的参数部分展开，从定义出发，重点结合用户主观评测实验对用户体验质量参数进行研究。实验部分进行的是用户主观评测实验，通过仪表模拟参数损伤制作视频片段，提供给一定数量的用户进行评测打分，对得到的评测结果数据进行验证、分析，最终获得用户体验质量的参数权重。另一方面，我们通过研究用户体验质量层参数和服务质量层参数之间的关联关系来进一步观察用户体验质量参数的意义。这部分主要通过数学理论建立模型来进行分析，将用户体验质量参数集整体引入到系统的设计中去。

3.3.2　非对称带宽的多通道自适应视频传输

本文根据网络的异构性和时变特性选择最新的标准作为研究对象，从提高适配能力着手对自适应视频传输技术进行深入的研究工作。

在视频领域，实现异构网络视频传输的传统方法是：在视频系统的主控单元（MCU，Master Control Unit）上针对不同的网络带宽和用户需求对原始大小视频流进行转码压缩。但由于转码压缩的复杂性，需要额外的硬件投资，成本很高。而且，随着异构环境的不断复杂化，这种转码的方法也将无法完全胜任。所以我们在异构网络的条件下，研究了采用自适应速率控制实现视频增强传输的方案。



RTP: Real-time Transport Protocol，实时传输协议
RTCP: Real-time Transport Control Protocol，实时传输控制协议
User Datagram Protocol，用户数据报协议

图10.　自适应视频流传输的层次结构

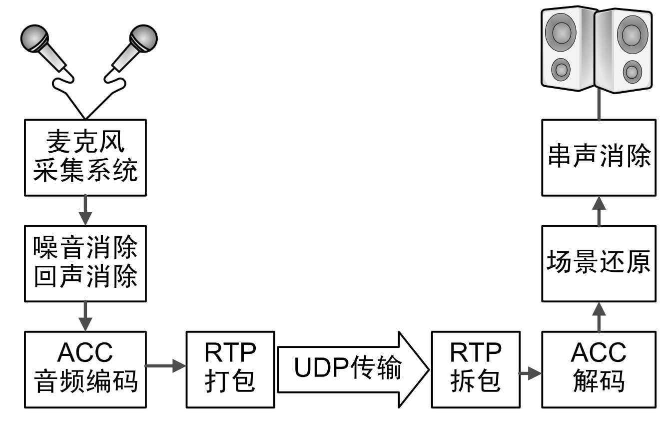
自适应速率控制通过使视频流速率与网络可用带宽匹配来避免网络拥塞的发生。速率自适应整形使发送端按照自适应速率控制指定的码率发送。网络的视频流自适应传输技术主要由 3 部分构成，即自适应速率控制、自适应视频和自适应速率整形，如图 10 所示。

我们首先进行源速率的控制，这样可以保持恒定的视频质量。同时考虑码流的传输系统。由于允许的传输速率是有限的，传输的帧间隔与视频帧的间隔是不同的。因此本技术根据对不同特征视频终端的分析，对码流的传输系统进行调整。由于允许的传输速率是有限的，传输的帧间隔与视频帧的间隔是不同的。为了使实时回放时具有更小的波动质量与更好的浏览效果，在实时视频通信系统中采用了公平性传输的方法，并结合视频通信系统具有多特征的特性，采用机会模型改进率平滑技术，提出基于特征的机会传输解决方法，通过控制视频流的速率等获得稳定的传输效果。

3.4 基于麦克风阵列的沉浸式高保真语音交互

我们在沉浸式高保真音频交互方面的研究内容是基于麦克风阵列，研究沉浸式高保真语音交互技术，并构建相应的系统。

采用的系统流程和系统主要功能模块如图 11 所示。事实上，远程交谈的每一端都既是音频输入端，又是音频输出端。在音频输入端，音频经麦克风阵列采集及噪音消除、回声消除等预处理后产生高质量的音频信号；单路音频信号编码后经由网络传输至远程音频输出端。在输出端，接收单路音频信号，进行场景还原，即采用 HRTF（head-related transfer function，头部相关传输函数）模型生成针对双耳的双声道音频，同时采用串声消除技术生成适合于扬声器直接播放的声音信号，并送至音箱播放。



ACC 音频编码：Advanced Audio Coding，一种基于 MPEG-2 的音频编码

图11.　沉浸式高保真音频交互系统流程

具体的工作为：

(1).　基于麦克风阵列的高保真音频采集和处理

当前已有的远程视音频互动系统多数面向视频会议的应用，在音频方面，一般为每位参与者单独配备麦克风，在参与者发言时开启麦克风以开始音频采集。这种方式可避免环境噪声，提高音频质量，但是操作相对繁琐，而且对用户有一定限制，要求用户必须靠近麦克风讲话。更为重要的是，这种方式丢失了说话人的位置、距离等信息，无法用于营造沉浸式的音频环境。因此，本文采用基于麦克风阵列的音频采集系统，能够采集较远距离(1~3 米)的声音，用户无需佩戴或手持麦克风即可自由参与交谈。

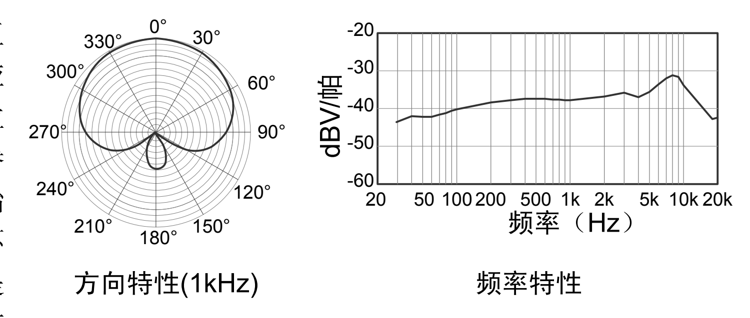


图12.　采用超心型指向麦克，保证采集的精度，降低场景噪音

在采集设备方面，我们的系统使用多个灵敏的超心型指向麦克风对远距离的用户语音进行捕捉。其频率和方向特性如图 12 所示。该麦克风具备较高的指向性，可保证采集的精度，降低场景噪音。

为了去除背景噪音，我们对使用麦克风阵列捕捉的各音频流信息进行自适应噪音消除，同时使用小波变换去除音频信号中的噪音，保留原始信号的主要成分；为了去除回声，我们利用一个自适应滤波器对未知的 LRM 回声[4]通道进行系统辨识，模拟回声路径，通过自适应滤波算法的调整，使其冲击响应与实际回声路径相逼近，从而得到回声预测信号，再将预测信号从麦克风接收到的语音信号中减去，从而实现回声抵消；为了提高音质，我们将先从音频信号中检测出人声，并通过对人声进行放大来提供音频增益。

(2). 音频场景建模和还原

音频场景建模和还原的目标是对人的双耳听觉效应进行建模，对于采集和传输来的音频信号进行场景还原，从而营造出具有方向和距离感的立体音频效果。

人的听觉之所以会产生立体感，主要取决于 ITD 和 IAD：

ITD（Inter Aural Time Delay）为两耳延迟的时间差。声波在空气中是以每秒 340m 的速度在传播，我们可以假设我们双耳之间的距离为 20cm，如果声源在右边，声音一定是先到达人的右耳然后再到达左耳，而这个延迟大概有 580μs。如果是正前方传来的，那么声音就会同时达到两个耳朵。所以很容易通过三角函数的方法得到声源所在方向。因此，人脑通过 ITD 可以毫不困难地得到声音的方位。

IAD（Inter Aural Amplitude Difference）为两耳音量大小差。当声音被物体挡住，所听到的声音会变小。如果声音从左方传来，那么左耳保留了原始音量，而右边的音量会减小，因为头部会吸收震动。所以说人也可以通过 IAD 来判断音源的位置。

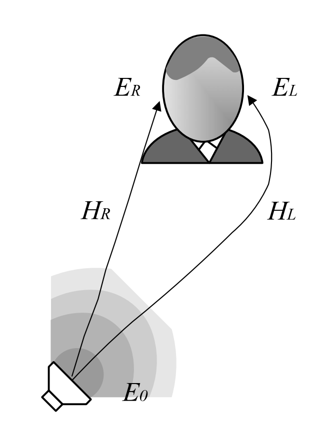


图13. HRTFs的左右耳图形表示

为进行立体声学场景的建模，本文系统采用了前述之 HRTF 函数技术。HRTF 是一组滤波器，系利用 ITD、IAD 和耳廓频率振动等技术产生立体音效，使声音传递至人耳内的耳廓、耳道和鼓膜时，聆听者会有环绕音效的感觉。通过数字信号处理，HRTF 可实时处理虚拟世界的音源。HRTF 参数可采用模拟声学实验获取。

使用 HRTF 数据实现虚拟立体声合成是指求 HRTF 的数据和预处理声音信号的卷积。若 H_R 和 H_L 分别代表右耳和左耳的 HRTF 冲击响应，$\boldsymbol{E_0}$ 代表输入，经过 HRTF 输出为：

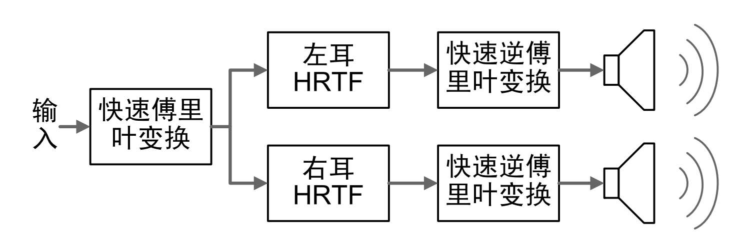


图14. 频域合成虚拟音频流程

$$\begin{cases} \boldsymbol{E_R} = \boldsymbol{H_R E_0} \\ \boldsymbol{E_L} = \boldsymbol{H_L E_0} \end{cases}$$

[4] 由扬声器—房间—话筒（Louder speaker-Room-Microphone）构成的声学回授所形成的回声

我们使用单声道信号，在频域内虚拟合成为具有临场感的音频。如图 14 所示，系统首先对声源信号进行快速傅里叶变换，变换后分别对左右声道进行 HRTF 计算，最后再通过快速逆傅里叶变换得到立体声音频。

由于我们的系统中采用了通过扬声器重放的方法，会引入交叉串声（Cross-talk），就是不但左耳，而且右耳也能听到左端扬声器发出的声音，反之亦然。这就破坏了双耳声道所还原的空间信息，因此双耳信号在重放前应该进行消除串声处理。

如图 15 所示，两个扬声器对称地放在收听者两侧（夹角为 2 θ），左右声道的双耳传输函数分别为 $\boldsymbol{H_{LL}}$，$\boldsymbol{H_{RL}}$，$\boldsymbol{H_{LR}}$，$\boldsymbol{H_{RR}}$，人的双耳听到的声压为：

$$\begin{bmatrix} P_L \\ P_R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \begin{bmatrix} L \\ R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} E_L \\ E_R \end{bmatrix} = [H][A] \begin{bmatrix} E_L \\ E_R \end{bmatrix}$$

其中 E_L,E_R 为 HRTF 计算后输出； A 为串声消除矩阵；L 与 R 为喇叭的输出

要实现串声消除，就需要通过计算得到合适的喇叭输出 $\boldsymbol{L}$ 和 $\boldsymbol{R}$，其关键在于确定串声消除矩阵 $\boldsymbol{A}$。选择串声消除矩阵的传输特性，使得$[\boldsymbol{A}]=[\boldsymbol{H}]^{-1}$，于是 $\boldsymbol{P_L}=\boldsymbol{E_L}$，$\boldsymbol{P_R}=\boldsymbol{E_R}$，这样扬声器的双耳声压与耳机的重放就相同了，从而消除了串声。一般扬声器的左右声道是对称的，即 $\boldsymbol{H_{LL}}=\boldsymbol{H_{RR}}=\alpha$，$\boldsymbol{H_{RL}}=\boldsymbol{H_{LR}}=\beta$，所以矩阵可以写成

$$[A] = \frac{1}{\alpha^2 - \beta^2} \begin{bmatrix} \alpha & -\beta \\ -\beta & \alpha \end{bmatrix}$$

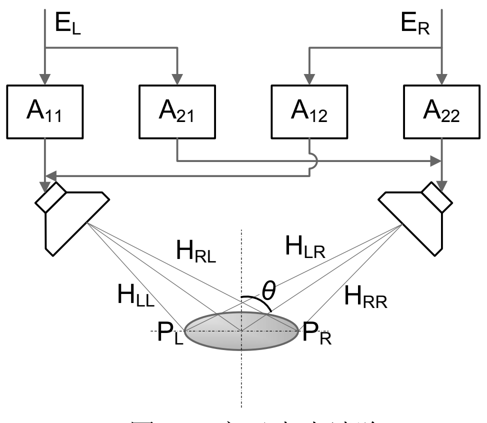


图15. 交叉串声消除

其中 α， β 为常数

综合基于 HRTF 的音频场景还原和串声消除，最终得到扬声器左、右声道的输出分别为：

$$L = \frac{\alpha H_L - \beta H_R}{\alpha^2 - \beta^2} E_0,\ R = \frac{\alpha H_R - \beta H_L}{\alpha^2 - \beta^2} E_0$$

其中 $\boldsymbol{H_R}$ 和 $\boldsymbol{H_L}$ 分别代表右耳和左耳的 HRTF 冲击响应，$\boldsymbol{E_0}$ 代表输入音频信号。

4 总结

为了提高远程视频自然交互的沉浸感，本文从四个方面做出了分析和研究，在服务于远程视频的音视频领域形成了知识产权壁垒。通过技术的突破，能够使异地交互的双方有同处一室的体验，能够保证实时的眼神交流和高保真的语音交互，同时能对网络传输的视频进行评估。随着网络电视的普及和发展，本文研究成果会有助于沉浸式技术走向产业化的道路。

作者简介：
黄美玉： 中国科学院计算技术研究所，普适计算研究中心，博士研究生，huangmeiyu@ict.ac.cn
尹苓琳： 中国科学院计算技术研究所，普适计算研究中心，硕士研究生
纪 雯： 中国科学院计算技术研究所，普适计算研究中心，副研究员
张博宁： 中国科学院计算技术研究所，普适计算研究中心，博士研究生
王向东： 中国科学院计算技术研究所，普适计算研究中心，高级工程师
陈益强： 中国科学院计算技术研究所，普适计算研究中心，研究员